Automatic Bayesian Induction of Classes

PETER CHEESEMAN, JIM KELLY, MATTHEW SELF AND JOHN STUTZ

AI RESEARCH BRANCH, MAIL STOP 244–17
NASA AMES RESEARCH CENTER
MOFFETT FIELD, CA 94035

NASA Ames Research Center
Artificial Intelligence Research Branch

Technical Report RIA–87–11–16–6

November, 1987

Conference Proceedings of The Second Annual Artificial Intelligence Research Forum, pp. 224-239. NASA AMES Research Center.

Automatic Bayesian Induction of Classes

Peter Cheeseman, Jim Kelly, Matthew Self and John Stutz

NASA Ames Research Center
Mail Stop 244-7
Moffett Field, CA 94035

November 4, 1987

Abstract

This paper describes a criterion, based on Bayes's theorem, that defines the optimal set of classes (a classification) for a given set of examples. This criterion does not require that the number of classes be specified in advance; this is determined by the data. Tutored learning and probabilistic prediction in particular cases are an important indirect result of optimal class discovery. Extensions to the basic class induction program include the ability to combine category and real valued data, hierarchical classes, independent classifications and deciding for each class which attributes are relevant.

1 Introduction

This paper describes a method for automatically discovering (inducing) classes from a given database. These classes can then be used to give insight into the patterns that occur in the particular domain, or make predictions in particular cases. This type of learning is often called unsupervised or untutored learning, since there are no preconceived classes and the number of classes to be found is not known. In supervised learning, on the other hand, the user expects the system to induce a specific classification based on a set of pre-classified examples. In either kind of learning, the resulting classification can be used to classify new cases. Many previous authors have published approaches in the area of automatic class discovery [9]. A large number of these approaches employ clustering methods which use a "similarity" measure that defines a "distance" between any pair of cases based on how "close" their descriptions are. Unfortunately, automatic clustering methods give different results depending on the similarity measure chosen. Even more disturbing is that automatic clustering methods require the user to specify the number of classes to be discovered, or rely on *ad hoc* methods for choosing an appropriate number of classes. As a result of the lack of a good criterion for choosing the number of classes, these methods often produce classes even if given randomly generated data. Consequently, the user never knows if the classes produced by clustering methods indicate actual classes in the domain or are just the result of random variation and the similarity measure used. Despite these criticisms, if there are strong natural classes in the data, clustering methods with any reasonable similarity measure will find them. It is when the natural classes are buried in excessive noise that clustering methods break

down.

The difficulties of clustering methods are well known, and have been strongly criticized in the AI learning literature—see [6],[9] for a clear critique. Michalski proposes an alternative method for automatic class discovery called "conceptual clustering". Briefly, this method uses a clustering procedure to produce potential classes, then uses a minimum description criterion (in a given language) to choose the final classes. The resulting classes clearly depend on the methods for generating candidate classes and the language used to describe them. Michalski regards this dependence as inevitable because he does not believe in "natural" classes. Instead, he believes that classes depend on the purpose of the classification, which in conceptual clustering is embedded in the class description language. Michalski's approach raises the fundamental question: in automatic classification, are classes *invented* or *discovered*? This is an old philosophical problem.

The approach described in this paper finds the *most probable* classification given the data. Since this is a Bayesian approach, prior probabilities for the number of possible classes need to be given, as well as prior probability distributions for all the parameters in the classification model. Intuitively, the most probable classification is the one in which the class descriptions (i.e. the probabilities of the attribute values) best predict the observed values. The most probable classification also decides the optimal number of classes as well as the class descriptions. Because it takes considerable data to move a class description away from its prior value, the Bayesian criterion is automatically biased against introducing more classes than the available data will support. The criterion for deciding which is the most probable classification is a direct consequence of Bayes's theorem (section 3), and does not require any *ad hoc* assumptions—in particular, it does not require any similarity measure, because it does not directly compare particular cases. The Bayesian criterion for optimal classsification essentially depends on how well the proposed classes predict the given data; the accuracy of the prediction could be regarded as a kind of similarity measure.[1] Although the classes found depend on the language used to describe the data, they do not depend on the "language" used to describe the classes, as in conceptual clustering.

The new automatic Bayesian class induction procedure allows both category-valued information (e.g. Sex) and real-valued information (e.g. Blood-pressure) in the data description. The particular method described in section 3.1 makes a number of assumptions. Section 4.2 discusses extended models that relax some of these assumptions. In particular, the assumption that all variables are important in describing all classes can be removed, as well as the assumption of mutual exclusivity of the classes. The Bayesian approach defines the criterion for optimal classification, but it does not say how to find the optimum—this is a difficult search problem. The use of the discovered classification for prediction purposes is discussed in section 5. This method of making predictions via a mapping onto classes is a common pattern of human reasoning, and corresponds closely to other supervised learning methods, such as ID3 [9]. The use of the class induction procedure for discovery of classes and their subsequent use for prediction can be viewed as automatic discovery of expert systems—without the expert.

[1] The Bayesian criterion is defined between the class descriptions and the cases, not between pairs of cases.

2 Data Format Definition

A typical data base of the form required is shown in Fig. 1. It assumes that all the data is in the form of an ordered (i.e. fixed) list of attributes that describe each case (example, item etc.). Attributes can either have category values, (e.g. Blood-Group), that have a predetermined set of possible values, or they can be real-valued, (e.g. Blood-Pressure). Catagory values are assumed to be mutually exclusive and exhaustive. Real-valued attributes are sometimes referred to here as variables. In the AI literature, reference is often made to "tree-valued" attributes. For example, the attribute SHAPE can be split into values [TRIANGLE, QUADRILATERAL, ··· etc.] at the first level, and the value TRIANGLE can be further subdivided into [REGULAR, ISOCELES, RIGHT-ANGLED, IRREGULAR] and the value IRREGULAR can be divided into [OBLIQUE, NON-OBLIQUE] and so on. However, the leaf nodes of the tree form a set of mutually exclusive and exhaustive values, so that the tree structure can always be flattened into a set of mutually exclusive and exhaustive values (as required by the above format). Tree structures may implicitly contained important prior probabilities information for the possible leaf values, leading to non-uniform prior probabilities. Such prior information should be extracted when the tree is flattened.

The data base format has a number of built-in assumptions that limit its generality. For example, it assumes that the number of possible outcomes for a given predicate is known (but the possibility of other values can be allowed for by adding an "other" catagory—see below). Although a possible attribute value is whether a patient is married or not, the format does conveniently represent the information that two particular patients are married to each other. Such relational information can be included by suitably extending the predicate. For the marriage example, a predicate "In-marriage" can be introduced, with specific values giving which marriage (if any) a particular individual is in. Married couples (or larger groupings) will all have the same marriage identifier. Another limitation is that the theory presented currently does not make use of ordering information. One form of ordering information that is ignored is orderings on catagory values of discrete attributes. For example, an attribute such as EDUCATION-LEVEL can have values such as [HIGH-SCHOOL-DROPOUT, HIGH-SCHOOL-GRADUATE, SOME-COLLEGE, etc.]—where an ordering on the values is understood. However, the theory presented below ignores such orderings by assuming that the only catagory labels matter. Similarly, the theory assumes that the order of the attributes that describe each case is arbitrary, so the results would be the same regardless of the order given. In other words, a data base is regarded as identical if the columns, such as in Fig. 1, are interchanged. When the attributes are particular wavelength intensities in the spectrum of an object, for example, this assumption is clearly ignoring important information.

If the data collector is not sure that the current catagories will cover all future values, a value "OTHER" can be added to allow for this possibility. However, the user must supply some prior probability that OTHER will occur or the system will give this probability a global default value. If the data base has missing data (indicated by *), this is treated as another possible value of the given attribute, which also must be given a suitable prior probability. The possible values of the "Sex" attribute, for example, are then: [Male, Female, *].

Cases	Blood Group	Sex	...	Human
Zaphrod Beeblebrox	O-	M	...	?
Peter Cheeseman	A+	M	...	Y
David Letterman	AB+	M	...	Y
Mickey Mouse	*	M	...	N
Minnie Mouse	*	F	...	N
⋮	⋮	⋮	⋮	⋮

Figure 1: Patient Data Base

3 Basic Bayesian Theory

We use Bayes's theorem to find the probability of each classification hypothesis H_J given all the data (cases) D. A classification hypothesis H_J is the hypothesis that there are exactly J classes, and is given by Bayes's theorem:

$$P(H_J \mid D) = c\, P(H_J) P(D \mid H_J) \tag{1}$$

where c is the normalizing constant (defined by the requirement that $\sum_{j=1}^{J} P(H_J \mid D) = 1$); and $P(H_J)$ is the prior probability that there are J classes. In the absence of any prior knowledge about the expected number of classes, we set all $P(H_J)$ priors equal, and so this term is absorbed into the normalizing constant c. If the user has some prior knowledge about the expected number of classes, this information can be inserted by choosing the appropriate non-equal priors $P(H_J)$. For example, the user might have prior knowledge that there are almost certainly classes present, so $P(H_1)$ is given very low weight. Existing programs that find classes in random data are essentially putting strong prior weight on the existence of classes, and so find classes in the noise. In Eqn. 1, D is the entire set of cases in the data base, as described in section 2. The only remaining undetermined term is the likelihood function $P(D \mid H_J)$. By definition, we have

$$\begin{aligned} P(D \mid H_J) &= \int\!\!\int P(D, \pi_j, \vec{\theta}_{jk} \mid H_J)\, d\pi_j d\vec{\theta}_{jk} \\ &= \int\!\!\int P(D \mid \pi_j, \vec{\theta}_{jk}, H_J)\, P(\pi_j, \vec{\theta}_{jk} \mid H_J)\, d\pi_j d\vec{\theta}_{jk} \end{aligned} \tag{2}$$

where π_j and $\vec{\theta}_{jk}$ are vectors of all the parameters that define the classification model, as defined below. In Eqn. 3, we integrate out (i.e. marginalize) the parameters that define the classes to get the simpler term $P(D \mid H_J)$ that is not a function of the class parameters. If the likelihood $P(D \mid \pi_j, \vec{\theta}_{jk}, H_J)$ (still a function of the class description parameters) is used instead of $P(D \mid H_J)$, as is usually done in maximum likelihood methods, it is not possible to make meaningful comparisons between hypotheses with different numbers of parameters. Maximum likelihood methods always [2] favor hypotheses with more parameters because they give more degrees of freedom to fit the likelihood

[2] Unless the number of parameters exceeds the amount of data.

to the data. In the case of classification, the "best" maximum likelihood fit occurs when there are as many classes as there are cases, and the class parameters match the case values exactly. The problem of overfitting when using maximum likelihood methods is well known, but the Bayesian solution to the problem is not, even though the Bayesian solution has been in the literature for nearly 50 years [8].

3.1 The Likelihood Function

The likelihood function, $P(D \mid \pi_j, \vec{\theta}_{jk}, H_J)$, is a measure of how likely the given data is to have come from the given class hypothesis H_J, defined by the parameters $\{\pi_j, \vec{\theta}_{jk}\}$. As an example, let us suppose our database consists of measurements of apples taken from an orchard. We have I apples and have measured K attributes of each (diameter, density, color, sugar content, etc.). The Bayesian classification method will produce a set of class descriptions based on these observed attributes that might correspond to J different varieties (Golden Delicious, Granny Smith, Pippin, etc.).

We use the following notation:

i	=	object number (1 to I)
j	=	class number (1 to J)
k	=	attribute number (1 to K)
x_{ik}	=	measured value of the kth attribute (diameter, density, etc.) of the ith apple
$\vec{x}_i$	=	vector of attributes of ith apple
$\{\vec{x}_i\}$	=	the set of attribute vectors of all I apples ($\equiv$ D)
$\vec{\theta}_{jk}$	=	vector of parameters which describe the probability distribution of the kth attribute of the jth class
π_j	=	class j probability—relative abundance of class j, e.g., the fraction of the sample of apples that are Granny Smiths

The likelihood represents the probability that the apples in our sample came from an orchard with a specified proportion of classes of apples (π_j), with each class defined by attribute value probabilities ($\vec{\theta}_{jk}$). To expand the likelihood further, we make the following assumptions. First, we assume the cases (examples, samples etc.) are independent, that is, the probability of getting our particular sampling of apples is equal to the product of the probabilities of getting each apple individually. This data independence assumption is equivalent to:

$$p\left(\{\vec{x}_i\} \mid \pi_j, \vec{\theta}_{jk}, H_J\right) = \prod_{i=1}^{I} p(\vec{x}_i \mid \pi_j, \vec{\theta}_{jk}, H_J)$$

Next, we assume the classes are mutually exclusive and exhaustive, so the probability of a single apple growing in our hypothetical orchard is equal to the sum of the probabilities that it came from each variety, weighted by the proportion of that variety in the orchard,

i.e.,

$$p(\vec{x}_i \mid \pi_j, \vec{\theta}_{jk}, H_J) \;=\; \sum_{j=1}^{J} \pi_j \, p(\vec{x}_i \in \text{class } j \mid \vec{\theta}_{jk}, H_J).$$

Finally, we assume all the attributes are (conditionally) independent within each class—i.e., given that an apple belongs to a particular class, the attributes within that class are independent of each other. This would not be the case if our measurements included both diameter and weight, since these quantities would show a strong correlation in every class. Under this assumption, we have:

$$p(\vec{x}_i \in \text{class } j \mid \vec{\theta}_{jk}, H_J) \;=\; \prod_{k=1}^{K} p(x_{ik} \mid \vec{x}_i \in \text{class } j, \vec{\theta}_{jk}, H_J).$$

The probability of getting a particular attribute value $x_{ik}$for a particular class j is a function of the definition of the class (as specified by the class parameters $\vec{\theta}_{jk}$) and the value (x_{ik}), i.e.,

$$p(x_{ik} \mid \vec{x}_i \in \text{class } j,\ \vec{\theta}_{jk}, H_J) \;=\; f(\vec{\theta}_{jk}, x_{ik}),$$

where we have assumed that the class parameters are independent of the number of classes (H_J). For example, the probability of a Granny Smith having a particular diameter is part of the definition of what constitutes a Granny Smith apple, and we must decide what sort of diameter distribution is appropriate for Granny Smiths. For real-valued variables, we will use a Gaussian probability distribution:

$$p(x_{ik} \mid \vec{x}_i \in \text{class } j,\ \vec{\theta}_{jk}, H_J) \;\approx\; \frac{1}{\sqrt{2\pi}\sigma_{jk}} \exp\left[-\frac{1}{2}\left(\frac{x_{ik} - \mu_{jk}}{\sigma_{jk}}\right)^2\right] \Delta x_{ik}$$

where Δx_{ik} is the error in measuring x_{ik}. In this equation, we are assuming that the measurement error is much less than the class standard deviation (i.e. $\Delta x_{ik} \ll \sigma_{jk}$). If this assumption holds, then the integral of the Gaussian curve for x_{ik} over the range Δx_{ik} is closely approximated by the value of the Gaussian at x_{ik}multiplied by its width Δx_{ik}, as given by the equation above. In choosing this equation to model our classes, we are describing the classes by two parameters, μ and σ, that together define the vector $\vec{\theta}_{jk}$.

$$\vec{\theta}_{jk} \;=\; \begin{bmatrix} \mu_{jk} \\ \sigma_{jk} \end{bmatrix}$$

This choice of a class model function may not be the best choice if, for instance, the harvesters have discarded apples whose diameters fall below a certain value. If this were the case, the correct probability distribution would be a truncated Gaussian. The Bayesian method is flexible with regard to the functions used to model the classes. Theoretically, any smooth, normalizable function can be used.

For catagory-valued attributes, such as apple color, we have

$$p(x_{ik} \mid \vec{x}_i \in \text{class } j,\ \vec{\theta}_{jk}, H_J) \ = \ p_{jkl}; \qquad \text{where } x_{ik} = l_k.$$

That is, for the discrete attribute k, the probability of the x_{ik}th value being in the category l_k is given by p_{jkl}. The set of probabilities p_{jkl} are the class parameters $\vec{\theta}_{jk}$in the discrete case.

Putting the above components together, the full likelihood function is

$$p\left(\{\vec{x}_i\} \mid \pi_j, \vec{\theta}_{jk}, H_J\right) \ = \ \prod_{i=1}^{I} \sum_{j=1}^{J} \pi_j \prod_{k=1}^{K} p(x_{ik} \mid \vec{x}_i \in \text{class } j,\ \vec{\theta}_{jk}, H_J).$$

This product can be expanded to give:

$$p\left(\{\vec{x}_i\} \mid \pi_j, \vec{\theta}_{jk}, H_J\right) =$$
$$\sum_{n_j's} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_J^{n_J} p(\{\vec{x}_i\}_{n_1} \mid j = 1) p(\{\vec{x}_i\}_{n_2} \mid j = 2) \cdots p(\{\vec{x}_i\}_{n_J} \mid j = J)$$

where $\{n_j\}$ represents a partition of all I cases into J classes with n_j cases in each class, and the summation is over all possible partitions of cases into classes. The terms in this sum are products of the probabilities of a particular partition of cases into classes, weighted by the product of powers of the class probabilities π_j. Note that $n_1 + n_2 + \cdots + n_J = I$. In general, this is sum has a very large number of terms.

3.2 Integration

The likelihood $p(\{\vec{x}_i\} \mid \pi_j, \vec{\theta}_{jk}, H_J)$, gives the probability of the data $\{\vec{x}_i\}$as a function of the class descriptions $\{\pi_j, \vec{\theta}_{jk}\}$. Using 1 above and suitable class priors $p(H_J)$, we can then calculate $p(H_J \mid \{\vec{x}_i\})$. By calculating this posterior for different values of J, we can then find the value of J that gives the most probable posterior value. That is, we can search for the class hypothesis H_J that gives the "best" number of classes. In order to be able to compare these posterior probabilities for different numbers of classes, we must integrate out the class parameters, as indicated in Eqn. ??. In more detail, we have:

$$\begin{aligned} p(\{\vec{x}_i\} \mid H_J) \ &= \ \int\int p(\{\vec{x}_i\} \mid \pi_j, \theta_{jk}, H_J) p(\pi_j, \theta_{jk} \mid H_J) d\pi_j d\theta_{jk} \qquad (3) \\ &= \ \int\int p(\{\vec{x}_i\} \mid \pi_j, \theta_{jk}) p(\pi_j, \theta_{jk}) d\pi_j d\theta_{jk} \end{aligned}$$

In the last step, we are assuming that the class parameters π_j and $\vec{\theta}_{jk}$ are independent of the number of classes J. Equation 4 is a multi-dimensional intergral over the full set of parameters. Note that Equation ??? already gives us the likelihood term, but the prior probabilities on the parameters $p(\pi_j, \theta_{jk})$ require further explanation. These prior probability distributions represent the user's knowledge of the possible values

of the parameters *before* the actual data is seen. Typically the prior probabilites on parameters are represented by an approximate range on the data.

This integral can be done in closed form for the real-valued and discrete valued cases. The intergal of the sum (Eqn. 3), becomes a sum of integrals; each integral having a closed form solution. These integrals are given explicitly in a companion paper due to lack of space. The result of this integration is a closed form solution to the probability of the data, $\{\vec{x}_i\}$, for each possible number of classes J. The problem is to find J with the maximum posterior probability. Unfortunately, for a large number of data points and many classes, the number of terms in the closed form solution is too large to evaluate exhaustively, so Monto Carlo sampling methods are being investigated to approximate this sum.

4 Assumptions

The assumptions (or limitations) built into the method described above are:

4.1 Assumptions

1. That classification is an appropriate model of the data. This is not true for temporal (non-static) data, for example.

2. That all attributes are useful in distinguishing classes—the extension to allowing every attribute to be either relevant or irrelevant to a particular class is discussed in the next section.

3. That all classes are independent of each other—this implies that knowledge of the probabilities of particular attributes in one class give no information about the underlying probabilities in any other class. One method of removing this assumption is to introduce hierarchical classes, discussed below.

4. Classifying all cases into a set of mutually exclusive and exhaustive classes is appropriate. Independent classifications, discussed below, provide an alternative.

5. That the attributes within a class are independent—i.e. the attributes are conditionally independent; conditioned on belonging to the given class. This is not correct when attributes such as Height, Weight, Length etc., are used, since they are all dependent on a common "shape" factor. It is possible to correct for such dependencies by using suitable joint probabilities, but these correction factors are not discussed here.

6. That all the data can be cast in the form of properties of individual cases—i.e., no relations between cases are permitted (see section 2).

7. That all class hypotheses (including ones with different numbers of classes) are equally likely a priori.

4.2 Relaxation of Assumptions

The Bayesian method does not care which model is used. By building different likelihood functions, the user can put in whatever model desired. However, in general, the more complex the model (i.e., the more adjustable parameters it contains), more data is needed to justify the additional complexity. We have considered the following models that relax some of the above assumptions.

4.2.1 Attribute Relevancy

The basic method presented above assumes that all the attributes are informative in deciding class membership. If most of the attributes are uninformative for a particular class, then the cost of estimating these parameters separately degrades the ability for the method to discover finer classes. This is because of the fundamental link between the number of parameters in a model relative to the amount of data. This restriction can be removed by specifying for each class which attributes are relevant to that class description. Those attributes that are relevant have their own class parameters to be estimated. Those attributes that are judged irrelevant to a particular class description are calculated using a single set of parameters that are shared across the corresponding classes. We note that different application domain differ considerably in the relevancy of the attribute. In the spectral classification experiments, the assumption that all attributes (spectral values) are relevant is justified, but for medical data bases we have investigated, where there are many very different attributes, most of which are not relevant to a particular class description.

4.2.2 Hierarchical Classes

The assumption that all classes are independent of each other (as well as being mutually exclusive and exhaustive) may not be correct in many applications. The independence assumption implies that knowledge of the probability distribution for attribute values in one class is non-informative about the corresponding distribution in another class. This assumption can be relaxed by introducing hierarchical classes, where classes closer together on the (hierarchical) tree are closer to each other. It is a simple matter to build classes hierarchically; the method is to form classes directly, then extract these class descriptions as data and recursively look for classes within the classes, and so on. Unless there is a huge amount of data, this process rarely goes beyond two levels. It is also possible to do hierarchical class splitting. The method is to first find the set of cases whose probability of belonging to the class to be split is suficiently high (e.g. 95%). The classification procedure is then applied to this subset of cases, using prior probabilities appropriate to this subset. The results when applied to the IRAS classes have revealed very interesting fine structure.

4.2.3 Independent Classifications

Perhaps the strongest assumption is that a classification is appropriate at all! That is, there are many situations where the assumption of mutually exclusive and exhaustive

classes is not appropriate. HIV infected patients are either [Non-symptomatic, Pre-AIDS, ARC or AIDS]—i.e., such patients *must* be one or other of these alternatives. On the other hand, diseases are not mutually exclusive, so it is possible for a patient to both have typhoid *and* cholera simultaneously. Given this situation, it is desirable to develop a classification model that allows multiple overlapping classifications. This means that a given case has a simultaneous probability distribution over each classification, and so can belong to more than one class, if the classes belong to different classifications. It is possible for a single case to have both AIDS and cholera. About the simplest model for multiple classifications is to assume that the classifications are completely independent of each other. This means that knowledge of the probability distribution in one classification says nothing about the distribution in another. The resulting likelihood of the ith value of the A attribute, given that it belongs to the jth class in the C classification *and* to the kth class in the D classification is:

$$p(A_i \mid C_j, D_k, \cdots) = \frac{p(A_i \mid C_j)p(A_i \mid D_k)\cdots}{p(A_i)^{n-1}} \tag{4}$$

where n is the number of classifications, and $p(A_i \mid C_j)$ is the probability of the ith attribute value given that the case is in the jth class under the Cth classification (and similarly for the Dth classification). The terms such as $p(A_i \mid C_j)$ are calculated as previously described in section 3.

4.2.4 Extended Models

The Bayesian criterion for finding the best classification model (and its various extensions discussed in this section) is of much greater generality. The derivation in section 3 has been specialized to classification, but other models are possible. For a description of applications of the Bayesian criterion to domains such as learning of grammars, finite state machines, line finding etc. see [7]. A natural extension of the automatic classification approach is to include time dependent data bases—i.e. trend analysis. This extension will require models of how systems evolve with time and priors on these possible models. Temporal models, such as Markov models and parameterized temporal models (e.g. Fourier, exponential decay etc.), are being investigated.

5 Prediction

AI and statistical pattern recognition literature often obscures the reason for finding good classifications. The classification work reported here was originally motivated by the desire to make (probabilistic) predictions directly from data. A previous approach [4],[5], based on maximum entropy, allowed the direct calculation of the (conditional) probability of any attribute value given any combination of other attribute values (i.e. given particular evidence). The domain information in the previous approach consisted of a set of joint or conditional probabilities (constraints) that summarized all the significant information about intervariable correlations in the domain. These significant constraints were found by comparing the *expected* probabilities of attribute value combinations with the *observed* values in a data base. Unfortunately, the cost of computing

RIA-87-11-16-6

Automatic Bayesian Induction of Classes

Peter Cheeseman, James Kelly, Matthew Self, and John Stutz — November 1987

This paper describes a criterion, based on Bayes' theorem, that defines the optimal set of classes (a classification) for a given set of examples. This criterion does not require that the number of classes be specified in advance; this is determined by the data. Tutored learning and probabilistic prediction in particular cases are an important indirect result of optimal class discovery. Extensions to the basic class induction program include the ability to combine category and real-valued data, hierarchical classes, independent classifications and deciding for each class which attributes are relevant.

RIA-88-02-01-01

Knowledge Servers – Applications of Artificial Intelligence to Advanced Space Information Systems

Peter Friedland — February 1988

We have begun a transition from passive information systems which act only to facilitate the storage and retrieval of stereotyped data to far more active and responsive systems which can deal with widely differing forms of human knowledge. Edward Feigenbaum has coined the term 'knowledge servers' to describe this next generation of active information management systems. Among the functions of a knowledge server will be: the ability to store enormous varieties of knowledge; the ability to determine, through natural discourse, the needs of its users; the ability to summarize and pursue complex relationships in its knowledge; the ability to test and critique user hypotheses and suggest previously unseen connections resulting from those hypotheses; and the ability to communicate and collaborate with other autonomous knowledge servers. Because of complexity and variety of information relevant to future major space missions like space station, these missions will act as a driving force and testbed for the knowledge server concept.

RIA-88-04-01-4

AutoClass: A Bayesian Classification System

Peter Cheeseman, James Kelly, Matthew Self, John Stutz, William Taylor, and Don Freeman — April 1988

This paper describes AutoClass II, a program for automatically discovering (inducing) classes from a database, based on a Bayesian statistical technique which automatically determines the most probable number of classes, their probabilistic descriptions, and the probability that each object is a member of each class. AutoClass has been tested on several large, real databases and has discovered previously unsuspected classes. There is no doubt that these classes represent new phenomena.

RIA-88-04-12-0

Learning by Making Models

Philip Laird — April 1988

We propose a theory of learning from unclassified data. The learning problem is that of finding the parameters of a stochastic process that best describes the incoming data stream. Special attention is given to the efficiency of the learning process, similar to Valiant's theory of supervised learning, and in contrast to conventional pattern recognition approaches. Illustrative domains are constructed and analyzed.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	Dates attached	

4. TITLE AND SUBTITLE

Titles/Authors - Attached

5. FUNDING NUMBERS

6. AUTHOR(S)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Code FIA - Artificial Intelligence Research Branch
Information Sciences Division

8. PERFORMING ORGANIZATION REPORT NUMBER

Attached

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Nasa/Ames Research Center

Moffett Field, CA. 94035-1000

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Available for Public Distribution

Peter Friedland 5/14/92 BRANCH CHIEF

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Abstracts ATTACHED

14. SUBJECT TERMS			15. NUMBER OF PAGES
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT